Journal of Biomolecular NMR **30:** 47–60, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

From NMR chemical shifts to amino acid types: Investigation of the predictive power carried by nuclei

Antoine Marin[a], Thérèse E. Malliavin[a], Pierre Nicolas[b,c] & Marc-André Delsuc[d]
[a]*Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique, 13 rue P. et M. Curie, 75 005 Paris, France;* [b]*Unité Mathématique Informatique et Génome, INRA – Domaine de Vilvert, 78 352 Jouy-en-Josas cedex, France;* [c]*Laboratoire Statistique et Génome, CNRS, Tour Evry2, 523 place des terrasses de l'Agora, F-91000 Evry, France;* [d]*Centre de Biochimie Structurale, CNRS UMR 5048 – INSERM UMR 554 – Université Montpellier-I, 29 rue de Navacelles, 34 090 Montpellier, France*

Received 12 February 2004; Accepted 4 May 2004

Key words: automatic assignment, Bayes theorem, Bayesian decision, BioMagResBank, chemical shift, NMR, structural proteomics

Abstract

An approach to automatic prediction of the amino acid type from NMR chemical shift values of its nuclei is presented here, in the frame of a model to calculate the probability of an amino acid type given the set of chemical shifts. The method relies on systematic use of all chemical shift values contained in the BioMagResBank (BMRB). Two programs were designed, one (BMRB stats) for extracting statistical chemical shift parameters from the BMRB and another one (RESCUE2) for computing the probabilities of each amino acid type, given a set of chemical shifts. The Bayesian prediction scheme presented here is compared to other methods already proposed: PROTYP (Grzesiek and Bax, *J. Biomol. NMR*, **3**, 185–204, 1993) RESCUE (Pons and Delsuc, *J. Biomol. NMR*, **15**, 15–26, 1999) and PLATON (Labudde et al., *J. Biomol. NMR*, **25**, 41–53, 2003) and is found to be more sensitive and more specific. Using this scheme, we tested various sets of nuclei. The two nuclei carrying the most information are C_β and H_β, in agreement with observations made in Grzesiek and Bax, 1993. Based on four nuclei: H_β, C_β, C_α and C', it is possible to increase correct predictions to a rate of more than 75%. Taking into account the correlations between the nuclei chemical shifts has only a slight impact on the percentage of correct predictions: indeed, the largest correlation coefficients display similar features on all amino acids.

Introduction

As the assignment of NMR spectra still requires manual expertise (Doetsch and Wagner, 1998), any method permitting to automate the assignment steps is useful in accelerating the spectral analysis in the frame of high throughput NMR structure determination (Montelione et al., 2000). The acquisition of NMR experiments permits to record two types of information: chemical shifts and correlations between spins. The correlations are widely used for spectral assignment and structure determination (Wüthrich, 1986; Ikura et al., 1990; Sattler et al., 1999), while chemical shift information is not used as systematically as correlations during the early stages of spectral analysis.

Nevertheless, some methods have been developed (PROTYP by Grzesiek and Bax, 1993, RESCUE by Pons and Delsuc, 1999 and PLATON by Labudde et al., 2003) for the amino acid type assignment from chemical shift values. On the other hand, chemical shift values can be used for determining secondary structures (Wishart and Sykes, 1994) or for determining the (ϕ,ψ) backbone angles (Cornilescu et al., 1999), once the sequential assignment has been performed. On a preliminary or refined structure, chemical shift quantification has been used for structure

*To whom correspondence should be addressed. E-mail: therese.malliavin@ibpc.fr docking (McCoy and Wyss, 2002), for structure quality evaluation (Williamson et al., 1995), for assignment of homologous proteins (Wishart et al., 1997), as well as for structure final refinement (Pearson et al., 1995).

We present here a way to calculate the probability of amino acid types from a set of chemical shift values through the use of the BioMagResBank, a databank of chemical shifts from assigned proteins (Seavey et al., 1991). The method is based on a probabilistic model and follows the work of Pons and Delsuc, 1999 (RESCUE) based on a neural network. Thus, although the underlying method is totally different, the method presented here has been implemented through a program called RESCUE2. This predictive method is used to determine the minimal set of nuclei giving the best predictions and to test various sets of nuclei available from NMR experiments.

This method has been developed for two purposes. First, it allows the systematic analysis of protein chemical shifts, in order to extract the maximum of chemical information in structure determination as well as in cases where the complete spectral assignment is not reachable. Secondly, we are here proposing and testing a very general scheme to process spectral information: this scheme can be used into a larger assignment strategy, as for instance in sequential assignment.

Theory

Bayesian decision scheme

The amino-acid type $u$takes its value among one of the 20 regular amino acid types and an amino acid of type u in the protein under study, is denoted AA_u. The set of chemical shift values observed for a given spin system is denoted X, and will be called an observation. The prior probability of an amino acid type is written $P(AA_u)$ and represents the probability of this amino acid type independent of the observation X.

We calculate the conditional probability $P(AA_u|X)$ of the amino-acid type u, given an observation X, using a probabilistic model of chemical shifts distributions. From the Bayes's theorem, one can write:

$$P(AA_u|X) = \frac{f_{AA_u}(X)P(AA_u)}{f(X)}, \tag{1}$$

where $f_{AA_u}(X)$ is the conditional probability of the set of chemical shift values X given the amino acid type u, and $f(X)$ is the sum of the $f_{AA_u}(X)$ terms over all amino acid types.

Since the quantity $f(X)$ is only a normalizing constant for the probabilities, it is not calculated. The prior probabilities $P(AA_u)$ of the amino acids are different among the AA_u, but, in the present work, in order not to bias the predictions in favor of the most probable amino acids, we assume the same prior probability ($P(AA_u) = 1/20 = 0.05$) for each amino acid. Therefore, Equation 1 simplifies to:

$$P(AA_u|X) \propto f_{AA_u}(X). \tag{2}$$

Nevertheless, the use of the present scheme for sequential assignment would necessitate to take into account the different $P(AA_u)$ values, in order to provide a prediction in agreement with the sequence composition. For this reason, our probabilistic model is formulated in the general context.

The computation of $f_{AA_u}(X)$ requires to relate the n observed chemical shift values to the $n_{\max}$ amino acid nuclei. Because of the overlapping of chemical shift ranges, there is no unique assignment of the observed chemical shift values to the nuclei of the considered amino acid. Therefore, the computation of $f_{AA_u}(X)$ must take into account all possible assignments, each one being incompatible with the others. We express this in the relation:

$$f_{AA_u}(X) = \sum_{\delta \in A(X,n)} f_{AA_u}(\delta), \tag{3}$$

where δ is a possible assignment of the chemical shifts to the amino acid AA_u nuclei, and $A(X,n)$ the set of all possible assignments of the n observed chemical shifts (X) to the $n_{\max}$ amino acid nuclei. If $n = n_{\max}$, the number of all assignments equals $n!$.

When the nuclei are not all observed in the spin system ($n < n_{\max}$), the sum in Equation 3 has to be calculated over the permutations of all combinations corresponding to a choice of n observed chemical shifts among the $n_{\max}$ spin system nuclei. This model includes the case of the nuclei having multiplicity larger than 1, as the H_β protons (or the H_α in Gly), whatever the number of observed chemical shifts. If a number of nuclei larger than the number present in AA_u is experimentally observed ($n > n_{\max}$), the probability $f_{AA_u}(X)$ is set to 0.

As the number of permutations of combinations may be too large to do all calculations even for a medium-sized amino acid, we developed an algorithm to generate permutations of combinations only for the most probable assignments. Practically, only assignments with probabilities $f_{AA_u}(\delta)$ greater than 10^{-6} were taken into account when computing the sum in Equation 3.

Independent chemical shifts model (ICSM)

Under the assumption that the observations of chemical shift values are independent inside a spin system, for a specific assignment of the chemical shift values to the spin system nuclei (δ), the probability $f_{AA_u}(\delta)$ can be written as a product over the n observed chemical shifts:

$$f_{AA_u}(\delta) = \prod_{j=1}^{n} f_{AA_u}(\delta_j), \tag{4}$$

where $f_{AA_u}(\delta_j)$ is the probability to observe the jth chemical shift with a value δ_j in the amino acid AA_u. Equation 4 is valid when the number n of chemical shifts observed is equal to the number $n_{\max}$ of nuclei in the amino acid.

For a sake of simplicity, we assume, as a first approximation, that the chemical shift distributions are Gaussian, and write Equation 4 as:

$$f_{AA_u}(\delta) = \prod_{j=1}^{n} \frac{\exp\left[-\frac{1}{2}\left(\frac{\delta_j-\mu_j}{\sigma_j}\right)^2\right]}{\sigma_j\sqrt{2\pi}}, \tag{5}$$

where μ_j is the mean value of the chemical shift j and σ_j its standard deviation.

As the number of observed chemical shifts may be smaller than the number of nuclei ($n \leq n_{\max}$), the probability of presence of the chemical shifts has to be taken into account. The probability $P(CS_j)$ of the presence of the jth chemical shift is determined from the analysis of missing chemical shifts in the BMRB assignments. In terms of this probability, Equation 5 can then be rewritten as:

$$\begin{aligned} f_{AA_u}(\delta) \;=\; & \prod_{i=1}^{n_{\max}} \left[1_{\{CS_i=NA\}}(1-P(CS_i))\right. \\ & \qquad \left. + 1_{\{CS_i\neq NA\}}P(CS_i)\right] \\ & \prod_{j=1}^{n} \frac{\exp\left[-\frac{1}{2}\left(\frac{\delta_j-\mu_j}{\sigma_j}\right)^2\right]}{\sigma_j\sqrt{2\pi}}, \end{aligned} \tag{6}$$

where $1_{\{CS_i=NA\}}$ equals 1 if the chemical shift i is not available (*NA*) and 0 otherwise. $1_{\{CS_i\neq NA\}}$ equals to $1-1_{\{CS_i=NA\}}$. Equation 6 has the advantage compared to Equation 4 that, if no chemical shift is observed ($n = 0$), a value for $f_{AA_u}(\delta)$ can be calculated, as the number $n_{\max}$ of nuclei in the spin system always differs from 0.

Correlated chemical shifts model (CCSM)

The hypothesis of independence of the chemical shift observations can be avoided by using multivariate normal distributions. In that frame, if $n = n_{\max}$, the density $f_{AA_u}(\delta)$ can be expressed as follows:

$$f_{AA_u}(\delta) = \frac{\exp\left[-\frac{1}{2}(\delta-\mathrm{M})^{\mathrm{T}}\Sigma^{-1}(\delta-\mathrm{M})\right]}{(2\pi)^{n/2}|\Sigma|^{1/2}}, \tag{7}$$

where δ is the vector of observed values $(\delta_1, \cdots, \delta_n)$, M the vector of the mean chemical shifts $(\mu_1, \cdots, \mu_n)$ for the nuclei, $(\delta-\mathrm{M})^{\mathrm{T}}$ T the transposed of the vector $(\delta-\mathrm{M})$, Σ the covariance matrix, Σ^{-1} its inverse, and $|\Sigma|$ its determinant. The term $(\delta-\mathrm{M})^{\mathrm{T}}\Sigma^{-1}(\delta-\mathrm{M})$ is known as the Mahalanobis distance (Mahalanobis, 1930).

The elements of the covariance matrix Σ are calculated as:

$$\Sigma_{ij} = \frac{1}{m-1}\sum_{k=1}^{m}(\delta_{ik}-\mu_i)(\delta_{jk}-\mu_j), \tag{8}$$

where m is the number of spin systems used to calculate the statistical parameters, and δ_{ik} (respectively δ_{jk}) the chemical shift of nucleus i (respectively j) in the spin system k. Equation 8 provides a non-biased estimator of the covariance because of the use of the factor $1/(m-1)$ instead of $1/m$ (Morgenthaler, 1997).

Using the density of Equation 7, and taking into account the probability of presence of chemical shifts (case $n \leq n_{\max}$), one can write the probability $f_{AA_u}(\delta)$ as:

$$\begin{aligned} f_{AA_u}(\delta) \;=\; & \prod_{i=1}^{n_{\max}} \left[1_{\{CS_i=NA\}}(1-P(CS_i))\right. \\ & \qquad \left. + 1_{\{CS_i\neq NA\}}P(CS_i)\right] \\ & \frac{\exp\left[-\frac{1}{2}(\delta-\mathrm{M}_u)^{\mathrm{T}}\Sigma^{-1}(\delta-\mathrm{M}_u)\right]}{(2\pi)^{n/2}|\Sigma_u|^{1/2}}, \end{aligned} \tag{9}$$

where M_u is the vector of mean values and Σ_u the covariance matrix for the amino acid AA_u.

Equation 9 is a generalization of the Equation 6. If we set the non-diagonal elements $\Sigma_{ij}(i \neq j)$ to 0 we recover the model assuming independence. In the frame of the two models presented here (ICSM and CCSM) and for $n \leq n_{\max}$, $f_{AA_u}(\delta)$ is given by Equations 6 and 9, respectively. Using these equations, it is possible to compute the probability of the 20 amino acid types given any set of chemical shifts observed in a spin system. The predicted amino acid type is chosen as the one with the highest probability.

Materials and methods

BMRB filtering

The complete BMRB (2371 NMR-STAR files) was downloaded in October 2002. Only protein assignments were kept (2300 files) and the following heuristic method was used to remove proteins with out-of-range chemical shift values; these proteins are essentially paramagnetic proteins. The mean chemical shift μ and the standard deviation σ of each nucleus were calculated within the BMRB. Then, if more than 5% of the chemical shift values for a protein, were located at more than $\pm 5\sigma$ from μ, the corresponding NMRSTAR file was removed. After the application of this heuristic filter, 2041 files remained.

Then, all sequences were aligned versus each other to remove homologous proteins that may introduce a bias into statistics and tests. As too small sequences lead to poor alignment statistics, the sequences containing less than 10 residues were also removed (1918 files left). The remaining sequences were then aligned using fasta3 (Pearson, 2000) and grouped using a 'maximum independent set of a bipartite graph' algorithm (Kashiwabara et al., 1992) with an E-value cut-off of 10^{-3}. Finally, in each of the 783 groups obtained, the sequence with the maximum number of chemical shifts available was selected within the following experimental conditions: pH in the 3–10 range and temperature in the 283–318 K range. The total filtering process is summarized in Figure 1. The complete list of non homologous BMRB entries used in this study can be downloaded from: ftp://ftp.cbs.cnrs.fr/pub/RESCUE2/BMRB_selected_sequences.txt

Test procedure

The predictions presented here have been calculated using the jackknife method (also known as leave-one-out cross validation) on the 783 non homologous sequences. The procedure was the following: one of the 783 sequences was removed from the list and the statistical parameters μ_i, σ_i, Σ_u and $P(CS_i)$ calculated on the remaining 782 proteins, are then used to predict the amino acid types in the removed sequence, using ICSM and CCSM. The procedure was applied to the 783 proteins, and the rate of correct prediction is the total overall efficiency obtained on all the spin systems of these proteins. In order to evaluate without bias the contribution of each nucleus to the prediction, only the spin systems containing exactly the set of nuclei on which the prediction is tested are processed. Predictions in case of missing chemical shifts are discussed in subsections 4.4 and 4.6. Depending on the nuclei studied, the number of predicted spin systems for each test lies between 20,000 and 45,000.

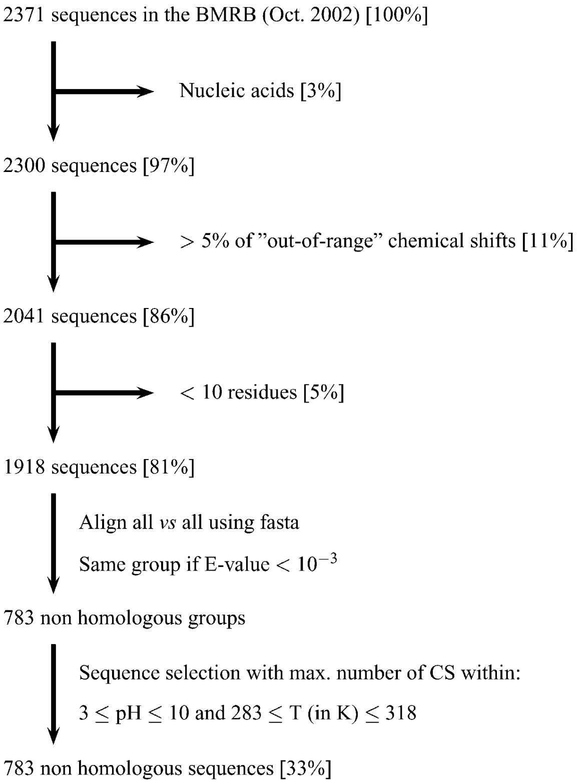
2371 sequences in the BMRB (Oct. 2002) [100%]
Nucleic acids [3%]
2300 sequences [97%]
> 5% of "out-of-range" chemical shifts [11%]
2041 sequences [86%]
< 10 residues [5%]
1918 sequences [81%]
Align all vs all using fasta
Same group if E-value < 10^{-3}
783 non homologous groups
Sequence selection with max. number of CS within:
3 ≤ pH ≤ 10 and 283 ≤ T (in K) ≤ 318
783 non homologous sequences [33%]

Figure 1. Protocol of the BMRB filtering to obtain a set of non-homologous proteins used in prediction tests.

Sensitivity and specificity

The quality of the prediction is analyzed by calculating the sensitivity and the specificity of the prediction for each amino acid type. For AA_u, the number of true positives (TP) is the number of correctly predicted spin systems of type AA_u, the number of false negatives (FN) is the number of spin systems of type AA_u, incorrectly predicted to other amino acid types, and the number of false positives (FP) is the number of spin systems of other types, incorrectly predicted to AA_u. The sensitivity of the method is the ratio $TP/(TP+FN)$ and the specificity the ratio $TP/(TP+FP)$, both ratios being expressed in percentage. If all the amino-acids are put in the same class (for a general efficiency estimation), only two classes remain (well predicted or badly predicted), and the specificity and sensitivity ratios are equal. In this latter case, we will use the term 'overall efficiency' to avoid confusions where these two ratios are different.

Software details

Two programs have been implemented: the first, called (BMRB stats) for computing the statistical parameters from the BMRB and the second, called (RESCUE2) for computing the probability of a spin system to correspond to each of the 20 amino acid types. The program BMRB stats takes as input a list of BMRB files from which the statistical parameters are extracted. The output is a file summarizing the nuclei statistics, including mean chemical shifts and standard deviation values, as well as the probabilities of presence and the covariance matrices. The program RESCUE2 takes as input the statistics computed using BMRB stats, and the list of spin systems to be analyzed. For each spin system, the output is the list of conditional probabilities $P(AA_u|X)$ of the 20 amino acids AA_u.

The programs have been implemented in the C language (ANSI), using the GNU Scientific Library* (GSL) (Galassi et al., 2002) for combination, permutation, vector and matrix operations. They are released under a free software** license (LGPL license), and are available at: ftp://ftp.cbs.cnrs.fr/pub/RESCUE2/. An online server is also available at: http://www.infobiosud.cnrs.fr/SERVEUR/RESCUE2/

If the total number of permutations of combinations were calculated, the computation time for the prediction with RESCUE2 on a Pentium® III at 500 MHz would vary from less than 0.1 second to more than months (time estimated for the most difficult case: the 13 hydrogens of Arg), depending on how many chemical shifts are used. A huge decrease is observed in the computation time, if only the most probable combinations and permutations are computed (as described above in subsection 2.1). For instance, in the case of Arg, the computation time decreases to few minutes. The average computation time for a full jackknife procedure on the test set is less than 3 h.

Results and discussion

Mean values, standard deviations and presences of chemical shifts in the BMRB

The mean chemical shifts and standard deviations calculated are similar to those given on the BMRB web site http://www.bmrb.wisc.edu/ref_info/statsel.html. The number of chemical shift observations is ranging from 3 (Hγ of Cys) to 5481 (H_N of Leu) with a mean value of about 1900 observations. The complete statistics can be found at: ftp://ftp.cbs.cnrs.fr/pub/RESCUE2/BMRB_stats.out

To be able to compare the different amino acids, we will focus on the following reduced set of nuclei: N, H_N, C_α, H_α, C_β, H_β and C'. The amino acids Pro and Gly will be ignored or treated as special cases since they do not possess all nuclei from the reduced set. The chemical shifts of nuclei having multiplicity larger than 1, as H_β protons (or H_α in Gly), are processed in the same way as the other chemical shifts.

The Gaussian distributions with the statistical parameters calculated on the BMRB (Figure 2) display different levels of overlapping among amino acids. Nuclei with very overlapping distributions and/or large standard deviations like H_N and C' would probably give few information about amino acid types. On the other hand, nuclei with dispersed distributions like H_β and C_β should give better information.

The percentages of missing chemical shifts are ranging from 3.4% (H_N of Ala) to 99.2% (C_ζ of Tyr) for the complete set of 274 nuclei. The most frequent nucleus observed is the H_N while frequently missing chemical shifts are generally exchangeable hydrogens and the long sidechain nuclei (C_δ to C_ζ, N_ε, N_ζ). For the 138 types of ^{1}H nuclei, 11 hydrogens were not measured for more than 75% of the cases, and were not considered further in the calculation. These hydrogens are the amide hydrogens of the Arg sidechain, the sulfur hydrogen of Cys, the exchangeable hydrogens of the His ring, the NH3+ terminal hydrogen of Lys, and the hydroxyl hydrogens of the Ser, Thr and Tyr sidechains. The least present hydrogen not excluded is the H_ε of Arg missing in 72% of the cases. In the reduced set of nuclei described at the beginning of this subsection, the sorted nuclei from the most present to the least present are: H_N (5.3%), H_α (14.9%), N (19.4%), H_β (22.8%), C_α (23.7%), C_β (30.0%) and C' (47.1%), where the percentages of missing chemical shifts are given in brackets.

[*] http://www.gnu.org/software/gsl/

[**] http://www.gnu.org

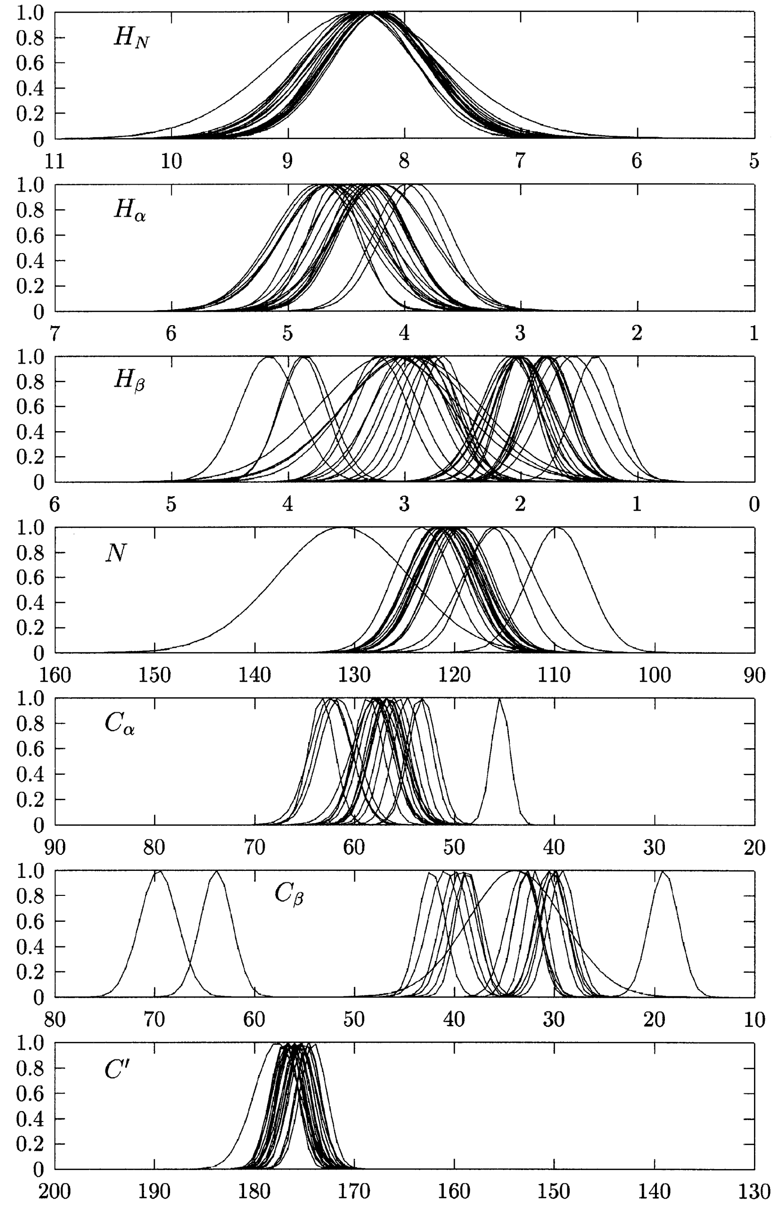

Figure 2. Gaussian distributions of the chemical shifts of the nuclei H_N, H_α, H_β, N, C_α, C_β, C′ among the 20 amino acids. The parameters of the Gaussian distributions were calculated from a statistical analysis of the 783 non homologous sequences of the BMRB.

Correlation coefficients between chemical shifts

The elements of the covariance matrix Σ_{ij} were normalized to correlation coefficients CC_{ij}:

$$CC_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}. \qquad (10)$$

The correlation coefficients (CC) concerning the reduced set of nuclei: N, H_N, C_α, H_α, C_β, H_β and C′, are displayed in Figure 3 with a palette of blue and red colors, corresponding to the intensity of the correlation. The tables containing the correlation coefficients are available as supplementary material. Slight differences are observed for amino acids of similar structures. Asp and Asn have similar correlation patterns, as well as (Glu, Gln), (Ile, Leu), and (Cys, Ser). Although the aromatic amino acids have the same chemical structure for the nuclei considered here, their different sidechains obviously influence in different

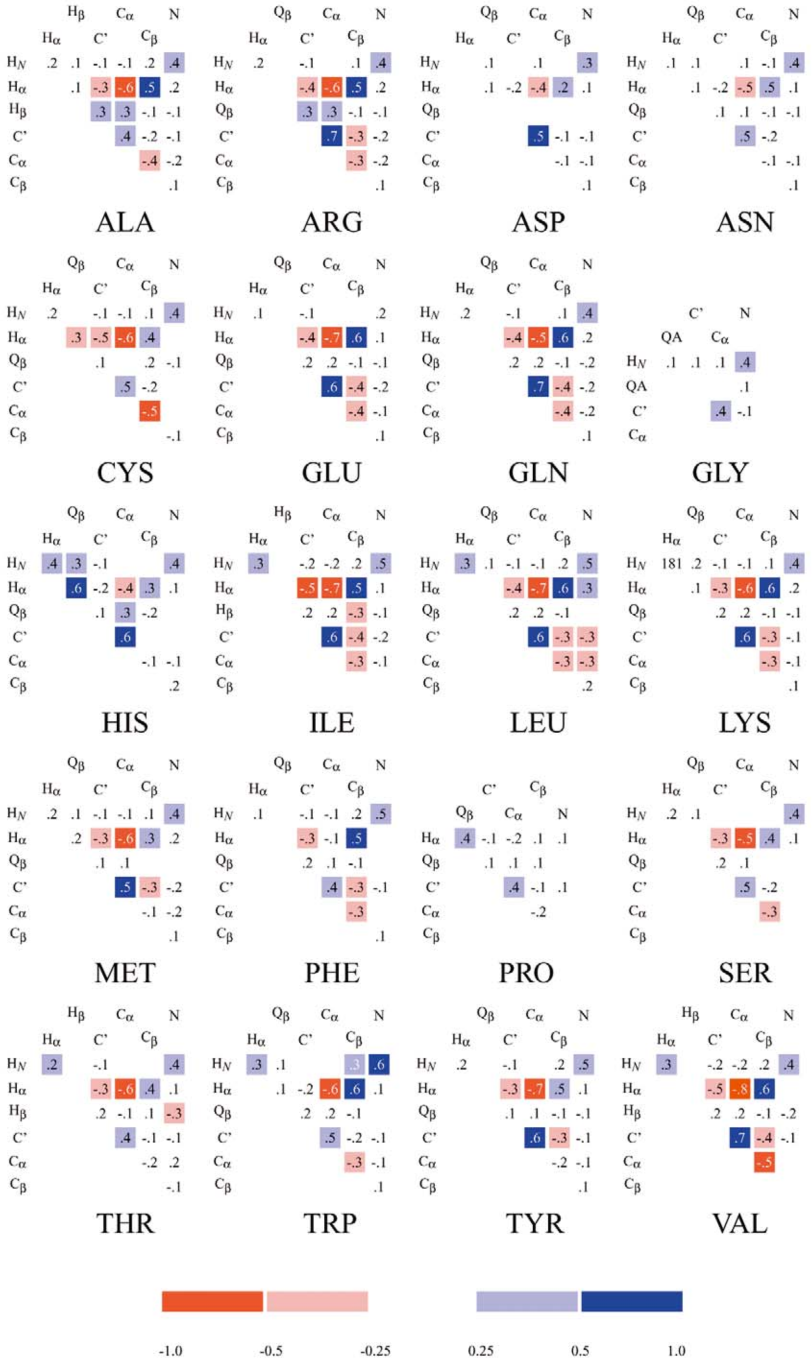

Figure 3. Correlation coefficients between the nuclei H_N, H_α (QA), H_β (QB), C', C_α, C_β and N for the 20 amino acids. The QB and QA spins are designing equivalent H_α and H_β hydrogens. The absolute values smaller than 0.25 are in white, the absolute values in the 0.25–0.5 range are in light color, and the absolute values larger than 0.5 are in dark color. For absolute values larger than 0.25, positive correlations are in blue, negative correlations in red. The correlation values, rounded off to the closest 1-digit number, are written if the absolute value is larger than 0.1.

ways the correlations between the nuclei close to the backbone. On the other hand, the amino acids (Ala, Arg) and (Lys, Leu) display similar patterns, despite of their different sidechains. A possible reason for this fact may be the distance of the sidechain groups to the backbone atoms. The amino acids Gly and Pro have correlations smaller than the other amino acids.

If only correlations larger than 0.5 (dark colors) are considered, the differences between the amino acids are less sensible. In that case, the amino acids Arg, Glu, Gln, Leu and Lys have exactly the same correlation pattern, and the amino acids Ala, Ile and Val display correlation patterns close to this one. The inclusion of the correlation coefficients is thus not expected to help much to discriminate between the amino acid types.

Mean values of the correlation coefficients have been calculated over all amino acids (except Pro and Gly). The average absolute value of the correlations is 0.21, which proves a strong correlation between the chemical shift values. There are 7 correlation larger than the average: H_α/C_α (−0.57), C_α/C' (0.55), H_α/C_β (0.48), H_N/N (0.42), H_α/C' (−0.34), C_α/C_β (−0.31) and H_α/H_N (0.22).

A strong negative correlation in the −0.75/−0.45 range is observed between the C_α and H_α chemical shifts, for all amino acids, except His (−0.38), Pro (−0.16) and Phe (−0.14). This strong negative correlation between the C_α and H_α chemical shifts is the basis of the Chemical Shift Index (CSI) method (Wishart and Sykes, 1994).

The strongest mean correlation values observed between the nuclei are determining three groups of nuclei: (N, H_N) , (H_α, C_β) and (C_α, C'), which are positively correlated. The strong negative correlations observed between H_α and C', H_α and C_α, C_α and C_β, are consistent with these groups. The smaller correlation between C' and C_β always exhibits a negative value in the −0.37/−0.076 range.

Predictive power of the different nuclei

The efficiency of the amino acid type prediction was tested extensively according to different combinations of nuclei. The best combinations are presented in Table 1 for the overall efficiency and in Table 2 for a summary of the sensitivities and specificities observed for the different amino acid types. The sensitivity ans the specificity ratios were defined above, in subsection 3.3. If all the amino-acids are put in the same class, only two predictions are possible (well predicted or badly predicted), the specificity and sensitivity ratios are equal, and the overall efficiency is the value of these ratios.

Table 1. Overall efficiency calculated for the ICSM and the CCSM for various sets of nuclei

Nuclei[a]	Independent CS model	Correlated CS model
H_N	7.4	–
C'	11.3	–
N	16.6	–
H_α	17.5	17.5
C_α	23.7	–
C_β	44.8	–
H_β	50.0	51.0
H_β, C_β	70.6	70.4
H_β, C_β, C_α	73.4	74.4
H_β, C_β, C_α, C'	73.0	76.2
All nuclei[b]	97.1	98.0

[a] The spin systems processed are those possessing all nuclei listed.
[b] Prediction made using all nuclei having a presence probability above 25%.

Single nuclei sets

ICSM was used for all tests made on single nuclei, H_β and H_α were also tested by CCSM. Surprisingly, the nucleus H_β gives the best overall efficiency (Table 1), but nucleus C_β gives more specific predictions (Table 2) for three amino acid types (Ala, Ser and Thr) versus two (Ser and Thr) for H_β. In the case of H_β, only four amino acids display a sensitivity smaller than 5% versus six amino acids for C_β. According to the efficiency of the prediction, the ranking of the nuclei is as follows: H_β, C_β, C_α, H_α, N, C' and H_N. The importance of the C_β chemical shift for the amino acid typing is in agreement with the observations already reported in the literature (Grzesiek and Bax, 1993).

The nuclei classification obtained here is in agreement with the qualitative observation of the chemical shift distributions in amino acids (Figure 2). Nuclei exhibiting smaller overlapping as H_β or C_β, are producing better prediction rates. Nevertheless, as the nucleus H_β displays more uniformly dispersed mean values compared to the nucleus C_β (Figure 2), the distribution of the mean values seems to influence the results of the prediction.

For all nuclei, except C_β and H_β, more than eight amino acid types display a sensitivity smaller than 5% (Table 2). Specificities larger than 95% are observed

Table 2. Amino acid types prediction summary for the CCSM results presented in Table 1

Nuclei	Spec. >95%[a]	Sens. <5%[c]
H_N	–	All except C, D, E, K
C'	–	All except A, G, L, N, P, T
N	–	All except A, G, K, L, N, P, S
H_α	G	All except D, E, G, I, N, P, V, W
C_α	G	C, F, H, M, Q, R, W, Y
C_β	A, S, T[b]	C, D, H, V, W, Y
H_β	S (92.9), T	H, M, R, Y
H_β, C_β	A, S, T, V (79.3)	H, Y
H_β, C_β, C_α	A, S, T, V (84.2)	C (14.6)
H_β, C_β, C_α, C'	A, L, S, T, V	C (13.6)
All nuclei	All except H, M, N, W	C (38.9)

[a] The sensitivity is specified after the amino acid name when <95%.
[b] Specificity >94.8% instead of >95%.
[c] If all amino acids were predicted with a sensitivity larger than 5%, the amino acid displaying the smallest sensitivity is shown with the sensitivity in brackets.

only for nuclei H_α, C_α, C_β and H_β (Table 2) and concern the amino acids Gly, Ala, Ser and Thr.

Best combined nuclei sets

After having tested each isolated nucleus, we determined the sets of two, three and four nuclei giving the highest prediction accuracy (Table 1). The set (H_β, C_β) displays the best overall efficiency (70%) as expected from the previous results. The addition of C_α adds 4% to the result obtained with (H_β, C_β). Finally, the overall efficiency reaches 76.2% with H_β, C_β, C_α and C'. The overall efficiency obtained on a set of nuclei is lower than the sum of the overall efficiencies on the single nuclei: there is some redundancy in the information carried by the chemical shifts. All nuclei sets show specificities larger than 95% for Ala, Ser and Thr (Table 2), in agreement with the observations made on single nuclei sets. On the other hand, the subsets of the three and four best nuclei do not have sensitivities smaller than 13% over the amino acids (Table 2). In both cases, Cys is the amino acid with the smallest sensitivity.

All nuclei

The efficiency of the amino acid prediction was tested using the complete set of chemical shifts available for ^1H, ^{15}N and ^{13}C nuclei in each amino acid with a probability of presence superior to 25%. This calculation is much too optimistic with respect of the usual possibilities of chemical shift measurements, as it is rare to measure the complete set of chemical shifts for all nuclei in a protein. But, the prediction performed here is intended to estimate the prediction power where almost all chemical shift information is available. The overall efficiency in this case is 98% using CCSM (Table 1). This result suggests that the complete set of chemical shifts may contain all information determining the amino acid type, which is not surprising for amino acids with a large number of nuclei, as these may be sufficient to discriminate between them. Nevertheless, few errors are observed for amino acids with a small number of nuclei except for Cys which is mostly predicted as Trp; it thus displays a sensitivity of 38.9% (Table 2).

The effect of correlations

A general feature of the predictions (Table 1) is the small increase (1–3%) observed for the overall efficiency between ICSM and CCSM. This is a consequence of the observation (Subsection 4.2) that large correlation values are observed between the same nuclei in all amino acid types. The largest difference (3.2%) is caused by the addition of the nucleus C' to the set of nuclei, in agreement with the large correlation between C_α and C'.

Nuclei sets from NMR experiments

RESCUE2 was also used to test different sets of chemical shifts corresponding to the recording of standard NMR experiments (Table 3).

Table 3. Overall efficiencies calculated in the scheme of the ICSM and CCSM, using different NMR experiments and their corresponding sets of chemical shifts

NMR experiment	Sets of chemical shifts	ICSM	CCSM
TOCSY (− aromatic)[a]	$H_N, H_\alpha, H_\beta, \ldots, H_\zeta$	78.5	79.9
TOCSY (+ aromatic)[b]	$H_N, H_\alpha, H_\beta, \ldots, H_\zeta$	86.5	88.3
^{15}N HSQC-TOCSY	H_N, H_α, H_β, N	52.4	53.4
HNCA	H_N, N, C_α	29.2	30.2
HNCA + HN(CA)CO	H_N, N, C_α, C′	33.5	39.7
HNCA + HN(CA)CO + CBCANH	H_N, N, C_α, C′, C_β	60.1	62.2
HNCA + HN(CA)CO + CBCANH + HNHA	H_α, H_N, N, C_α, C′, C_β	63.9	70.5

[a] Results obtained without the aromatic nuclei.
[b] Results obtained using the aromatic nuclei having a presence probability greater than 25%.

First, ^{1}H chemical shifts were used, as they can be measured in TOCSY experiments. Sufficient magnetization transfer was assumed, in order to observe hydrogens in the sidechains farther than the H_β. Two tests have been performed, the first one where aromatic hydrogens were not included in the calculation, and the second one where aromatic spin systems were assumed to be correctly connected to the backbone spin systems.

A second set of chemical shifts was used, according to the acquisition of a ^{15}N HSQC-TOCSY experiment. Only the N, H_N, H_α and H_β chemical shifts were selected. This assumption is valid for a protein mass smaller than 10 kD.

Then different sets of triple resonance experiments were tested: HNCA to measure the H_N, N and C_α chemical shifts, HNCA and HN(CA)CO to measure the H_N, N, C_α and C′ chemical shifts, HNCA, HN(CA)CO and CBCANH (or HNCACB) to measure the H_N, N, C_α, C′ and C_β chemical shifts. The set of HNCA, HN(CA)CO, CBCANH and HNHA experiments to measure the H_N, N, C_α, C′, H_α and C_β chemical shifts, was also tested.

The results presented here have been obtained using nuclei with a probability of presence superior to 25%. This procedure is intended to give a non-biased estimation of the prediction efficiency. The obtained rates of correct predictions are thus larger than the ones expected using spin systems with missing chemical shifts, and can be considered as upper bounds of the prediction efficiency. On the other hand, lower bounds of the efficiency can be estimated from the tests on single nuclei (Subsection 4.3).

The tests were made for the nuclei contained in the considered amino acid types: for instance, the prediction for the Glycine did not use the C_β chemical shift. Also, the Proline was excluded from all tests except for the TOCSY experiment.

TOCSY

The prediction of the amino acid type based on the ^{1}H chemical shift values without aromatic chemical shifts (Table 3) is displaying sensitivity and specificity values above 75% for the majority of the 20 amino acids (Ala, Arg, Val, Gln, Glu, Gly, Leu, Lys, Pro, Ser, Thr). The amino acids Cys, His, Ile, Trp and Tyr, which are mainly aromatic or of AMX topology, have both specificity and sensitivity factors smaller than 75%. These smaller values are due to prediction errors among AMX amino acids.

The use of aromatic chemical shifts increases the overall efficiency of about 10% (Table 3). This increase is due to a better prediction of the aromatic amino acids, which is not surprising, because of the specific information included in the input. Furthermore, a better prediction of some AMX spin systems (Cys, Asp) is achieved by avoiding some miss-predictions.

^{15}N TOCSY-HSQC

The prediction based on N, H_N, H_α and H_β chemical shifts shows smaller overall efficiency (Table 3) than the previous prediction because many ^{1}H signals have been removed from the observations. Three amino acids (Ala, Gly and Ser) display sensitivity and specificity values larger than 75%: among them, Ala and Ser already displayed large sensitivity for the following sets of nuclei: H_β, C_β, (H_β, C_β), (H_β, C_β, C_α) and (H_β, C_β, C_α, C′) (Table 1).

Triple resonance spectroscopy

Using the HNCA experiment (Table 3), only 30.2% of the spin systems can be predicted correctly. Upon

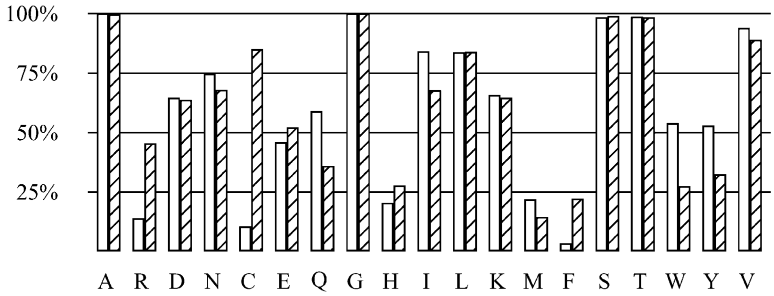

Figure 4. Sensitivity (solid) and specificity (hatched) results for all amino acid types in the frame of the CCSM using the H_α, H_N, N, C_α, C' and C_β nuclei. The number of predicted spin systems in this test is 26,148.

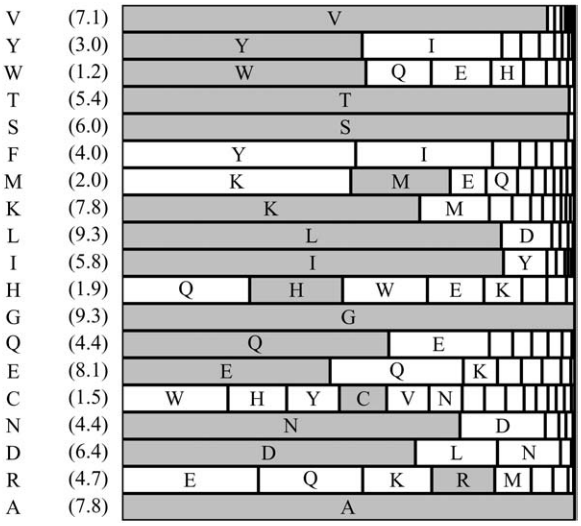

Figure 5. Prediction results for CCSM using the H_α, H_N, N, C_α, C' and C_β nuclei. The types of amino acid predicted are displayed according to the type of the amino acid in the prediction input (left of the figure). The prediction results are given for CCSM using the H_α, H_N, N, C_α, C' and C_β nuclei. The numbers given in brackets are the population sizes of the input in percentage. For each amino acid type, the proportion observed for its own type (in gray) represents the sensitivity for this type.

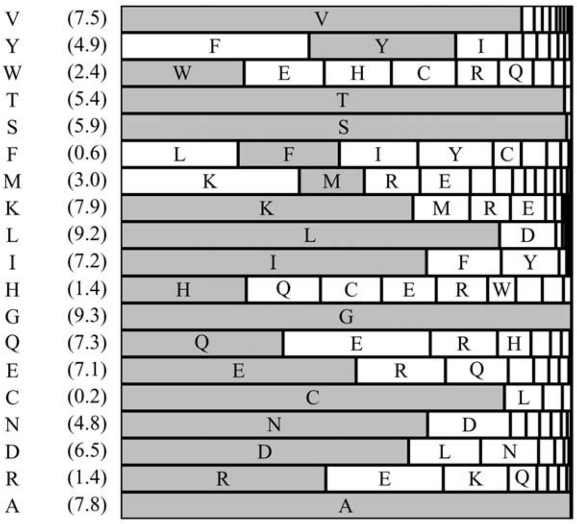

Figure 6. Prediction results for the CCSM using the H_α, H_N, N, C_α, C' and C_β nuclei. The types of the amino acids in the prediction input are displayed according to the predicted type of amino acid (shown at the left of the figure). The numbers given in brackets are the population sizes of the prediction output in percentage. For each amino acid type, the proportion observed for its own type (in gray) represents the specificity for this type.

addition of the C' chemical shift measured in the HN(CA)CO experiment, the overall efficiency reaches 39.7%. As expected, the addition of the C_β clearly improves the prediction, increasing the overall efficiency up to 62.2%. With the addition of H_α, the final result (70.5%) does not differ substantially from the result obtained with the (C_β, H_β) subset (Table 1). Differences in the efficiency between the ICSM and the CCSM larger than 6% can be observed after adding the C' and H_α chemical shifts. This may be due to the large correlations involving H_α and C' (Subsection 4.2).

We analyze in more details the last example (nuclei H_N, N, C_α, C', C_β, H_α) in the frame of CCSM, by showing the sensitivity and specificity for each amino acid type (Figure 4), the predictions obtained for each amino acid type proposed in input (Figure 5), and the amino acid type of the input sorted according to the amino acid type predicted (Figure 6).

The majority of the amino acids displays similar values of sensitivity and specificity (Figure 4). Two major kind of exceptions are observed. The first one is shown by the Cys (Arg displays similar behavior); its sensitivity is small (10.4%) but its specificity is high (85.1%). The Cys input spin systems are predicted (Figure 5, line 15) as several other amino acid types, in particular Trp, His, Tyr, Val, Asn, which produces the small sensitivity. On the other hand, due

Table 4. Compared prediction results for RESCUE and RESCUE2 (ICSM) using the ^{1}H chemical shifts and the ^{15}N HSQC-TOCSY chemical shifts

Amino Acid	RESCUE ^{1}H	RESCUE ^{15}N	RESCUE2 ^{1}H	RESCUE2 ^{15}N
A	92.2	70.2	95.5	89.9
C	23.7	17.7	17.4	10.7
D	48.3	40.5	85.5	76.1
E	39.8	24.7	93.7	30.4
F	8	1.3	5.1	6.2
G	91.2	79.7	88.1	87.2
H	1.9	15.8	0.8	2.1
I	82.7	43.7	73.8	51.1
K	92.7	1.3	100.0	56.3
L	64.9	57.0	79.5	46.5
M	50.7	8.2	89.3	5.2
N	10.4	29.1	97.9	29.3
P	77	–	90.7	–
Q	51.8	17.7	100.0	51.1
R	90.3	27.8	99.1	6.2
S	89	60.8	90.7	71.1
T	90.8	72.8	99.8	74.9
V	94.5	50.0	95.2	57.0
W	60	27.2	65.3	69.1
Y	0	17.7	20.0	14.1
All	63.5	34.9	79.9	53.4

to the high specificity, the majority of predicted Cys are effectively Cys spins systems (Figure 6, line 15). Consequently, although Cys is 1.5% of the total number of tested amino acids, the predicted Cys represents only 0.2% of the predictions. The second case is represented by the Trp (but Tyr, Gln and Ile have similar features); its sensitivity (53.9%) is greater than its specificity (27.4%). Indeed, while half of the Trp is predicted as Trp, the other half is mainly predicted as Gln, Glu and His (Figure 5, line 3). But, due to the small specificity, a significant amount of Glu, His, Cys, Arg and Gln is predicted as Trp (Figure 5, line 3), with the consequence that the number of predicted Trp spin systems equals twice the number of the input Trp spin systems.

Comparison with rescue

The predictions obtained with RESCUE2 (ICSM) were compared (Table 4) with results obtained with RESCUE on the ^{1}H chemical shifts (Pons and Delsuc, 1999), and on the N, H_N, H_α, H_β chemical shifts (Auguin et al., 2003). The prediction based on ^{1}H chemical shifts with RESCUE2 gives significantly better results than with RESCUE (Pons and Delsuc, 1999). Indeed, only nine amino acids (Gly, Ala, Val, Ile, Pro, Thr, Lys, Arg, Ser) display a sensitivity better than 75% using RESCUE, whereas 16 amino acids (Ala, Arg, Asp, Asn, Gln, Glu, Gly, Leu, Lys, Met, Pro, Ser, Thr, Trp, Tyr, Val) display a sensitivity better than 75% using RESCUE2. Isoleucine is the only case where the prediction is slightly worse, however, Ile has a sensitivity value of 71.2% and a specificity value of 66.9%. While the overall efficiency for RESCUE is 63.5%, we observe a value of 79.9% for RESCUE2.

The prediction with RESCUE2 using N, H_N, H_α, H_β chemical shifts is also displaying better results (Table 4) than those obtained with RESCUE (Auguin et al., 2003). Indeed, nine amino acids (Ala, Asp, Gly, Ile, Leu, Pro, Ser, Thr, Val) show sensitivity larger than 50%, compared to only five amino acids (Ala, Gly, Leu, Thr, Ser) in the RESCUE analysis. The overall efficiency is 53% for RESCUE2, compared to 35% in the RESCUE analysis.

It is not obvious to tell why the probabilistic model of RESCUE2 has a better performance than the neural network approach used in RESCUE, but the main differences of RESCUE2 in comparison to RESCUE are:

1. the direct use of observed values without a fuzzy logic input layer;
2. the inclusion of all possible assignments of chemical shifts to the nuclei, by taking into account the permutations of combinations in Equation 3;
3. a larger and a less redundant learning set;
4. the use of the probabilities of presence.

Comparison with PROTYP and PLATON

RESCUE2 was compared with the PLATON and PROTYP methods presented in Labudde et al., 2003. These authors used the C_α, H_α, C_β and C' chemical shifts recorded for a set of 51 proteins. The prediction was performed with RESCUE2 using the same data sets for learning and prediction (Table 5, RESCUE2 (a)). The overall efficiency of the prediction was analyzed for the first ranked amino acid (1st), and for the three first ranked amino acids (1–3). RESCUE2 is giving better results than PLATON and PROTYP, with a difference of overall efficiency always greater than 10%.

The predictions performed with PLATON only allowed the use of complete spin systems. An additional

Table 5. Comparison of the prediction results obtained with RESCUE2 and PLATON/PROTYP. The results were analyzed like in Labudde et al. (2003) a prediction is successful either if the first ranked prediction is the correct amino acid (1st) or if the correct prediction is in the three first ranked prediction (1–3). PLATON(1) and PLATON(2) correspond to the different penalty functions used in Labudde et al., 2003. RESCUE2(a) and RESCUE2(b) correspond respectively to the predictions without and with processing of incomplete spin systems

Rank	PLATON(1)	PLATON(2)	PROTYP	RESCUE2(a)	RESCUE2(b)
1st	45.4	61.7	59.3	72.5	71.5
1–3	72.0	83.1	86.8	95.2	94.2

prediction was run with RESCUE2, taking into account incomplete spin systems (Table 5, RESCUE2 (b)), through the probabilities of presence. The results of this second prediction show only a slight decrease (about 1%) in the overall efficiency, and the number of predicted spin systems increased by 9%.

Conclusion

A probabilistic approach was presented for the prediction of amino acid type from the chemical shift values of a given spin system. This approach is based on an extensive use of the chemical shift values recorded in the BMRB. A protocol and a set of programs were developed to allow the automatic processing of any BMRB subset.

Several authors (Gronwald et al., 1998; Cornilescu et al., 1999) pointed out the problem of the correct referencing of the chemical shifts values, specially in the case of ^{13}C nuclei. This referencing is required for any quantitative use of these values. Here, we deliberately ignored this aspect, and processed all chemical shift values stored in the BMRB for two reasons: (i) to rely on the maximum-size data bank, and (ii) the difficulty to sort the incorrectly referenced protein assignments from the correctly referenced ones. The results presented here may have suffered from such an approximation but they are still improving the prediction achieved by other methods.

Indeed, the probabilistic scheme proposed here shows more sensitive and specific results than all other previously proposed approaches (Grzesiek and Bax, 1993; Pons and Delsuc, 1999; Auguin et al., 2003; Labudde et al., 2003). Progress were not only made in the efficiency of the prediction but also in the modeling of the assignment problem, by taking into account the missing chemical shifts and the different possible assignments of chemical shifts to nuclei. The improvements in the modeling permit to use RESCUE2 on any set of chemical shifts, and would be of major importance in a larger assignment strategy.

The RESCUE2 input is only specifying the type of nuclei for which chemical shifts are measured, i.e., ^{15}N, ^{13}C or ^{1}H. Inside the same type of nucleus, the chemical shifts are provided without specific atom indication. This is an important advantage of the method in automatic processing of the chemical shifts, and is made possible by the algorithm designed to generate permutations of combinations only for the most probable assignments (Subsection 2.1).

The efficiency of the various nuclei to predict the amino acid type has been extensively investigated. The H_β was identified as the most efficient nucleus, followed by the nuclei C_β, C_α, H_α, N, C′ and H_N. We can achieve prediction rates as high as 51% with only one nucleus (H_β), 70.6% for two nuclei (H_β, C_β), 74.4% for three nuclei (H_β, C_β, C_α) and 76.2% with four nuclei (H_β, C_β, C_α, C′). Finally, we have been able to show that from a typical set of NMR experiments, it is possible to reach an overall prediction efficiency as high as 70%.

The first practical conclusion of this work is that the chemical shifts of an isolated spin system, recorded in a set of usual triple resonance experiments, contain sufficient information to reach a prediction level up to 70 %, without using information about the adjacent residues. This can be very useful in the case of interaction studies between biomolecules, for which the sequential assignment is unreachable (Ramaen et al., 2003). Another conclusion is that the chemical shifts of two nuclei, H_β and C_β, are sufficient to reach an overall efficiency higher than 70%. Finally, the correlations between the chemical shifts are large, but not sufficiently specific to the amino acid to improve the prediction using multivariate distributions.

Acknowledgements

Antoine Marin thanks the 'Ministère de la Recherche et de la Technologie' and INRIA for the postdoctoral financial support in the frame of the program 'Bioinformatique' (project 'Géométrie de Distance et Génomique Structurale à haut débit par RMN'). CNRS, INSERM, University Paris-VII and University Montpellier-I are acknowledged for funding.

Supporting Information Available: Tables containing the correlation coefficients for each amino-acid (PDF). This material is available free of charge via the internet at: ftp://ftp.cbs.cnrs. fr/pub/RESCUE2/ supplem_corr.pdf and http://kluweronline.com/issn/ 0925-2738.

References

Auguin, D., Catherinot, V., Malliavin, T., Pons, J.-L. and Delsuc, M.-A. (2003) *Spectroscopy*, **17**, 559–568.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.

Doetsch, V. and Wagner, G. (1998) *Curr. Opin. Struct. Biol.*, **8**, 619–623.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M. and Rossi, F. (2002) *GNU Scientific Library Reference Manual*, ISBN 095416170X.

Gronwald, W., Willard, L., Jellard, T., Boyko, R., Rajarathnam, K., Wishart, D., Sonnichsen, F. and Sykes, B. (1998) *J. Biomol. NMR*, **12**, 395–405.

Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.

Ikura, M., Kay, L. and Bax, A. (1990) *Biochemistry*, **29**, 4659–4667.

Kashiwabara, T., Masuda, S., Nakajima, K. and Fujisawa, T. (1992) *J. Algorithms*, **13**, 161–174.

Labudde, D., Leitner, D., Kruger, M. and Oschkinat, H. (2003) *J. Biomol. NMR*, **25**, 41–53.

Mahalanobis, P. (1930) *J. Asiatic Soc. Benagal*, **26**, 541.

McCoy, M. and Wyss, D. (2002) *J. Am. Chem. Soc.*, **124**, 11758–11763.

Montelione, G., Zheng, D., Huang, Y., Gunsalus, K. and Szyperski, T. (2000) *Nat. Struct. Biol.*, **7**, 982–985.

Morgenthaler, S. (1997) *Introduction à la statistique*, Presses polytechniques et universitaires romandes.

Pearson, J., Wang, J., Markley, J., Le, H. and Oldfield, E. (1995) *J. Am. Chem. Soc.*, **117**, 8823–8829.

Pearson, W. (2000) *Meth. Mol. Biol.*, **132**, 185–219.

Pons, J.-L. and Delsuc, M.-A. (1999) *J. Biomol. NMR*, **15**, 15–26.

Ramaen, O., Masscheleyn, S., Duffieux, F., Pamlard, O., Oberkampf, M., Lallemand, J.-Y., Stoven, V. and Jacquet, E. (2003) *Biochem. J.*, **376**, 749–756.

Sattler, M., Schleucher, J. and Griesinger, C. (1999) *Prog. NMR Spectrosc.*, **34**, 93–158.

Seavey, B., Farr, E., Westler, W. and Markley, J. (1991) *J. Biomol. NMR*, **1**, 217–236.

Williamson, M., Kikuchi, J. and Asakura, T. (1995) *J. Mol. Biol.*, **247**, 541–546.

Wishart, D. and Sykes, B. (1994) *J. Biomol. NMR*, **4**, 171–180.

Wishart, D., Watson, M., Boyko, R. and Sykes, B. (1997) *J. Biomol. NMR*, **10**, 329–336.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley-Interscience.